8.3 Caches

Cache: a safe place for hiding or storing things.

Webster's New World Dictionary of the American Language, Second College Edition (1976)

Cache is the name first chosen to represent the level of the memory hierarchy between the CPU and main memory, and that is the dominant use of the term. While the concept of caches is younger than the IBM 360 architecture, caches appear today in every class of computer and in some computers more than once. In fact, the word has become so popular that it has replaced "buffer" in many computer-science circles.

The general terms defined in the prior section can be used for caches, although the word *line* is often used instead of block. Figure 8.5 shows the typical range of memory-hierarchy parameters for caches.

Block (line) size	4 – 128 bytes
Hit time	1 – 4 clock cycles (normally 1)
Miss penalty	8 – 32 clock cycles
(Access time)	(6 – 10 clock cycles)
(Transfer time)	(2 – 22 clock cycles)
Miss rate	1% – 20%
Cache size	1 KB – 256 KB

FIGURE 8.5 Typical values of key memory-hierarchy parameters for caches in 1990 workstations and minicomputers.

Now let's examine caches in more detail by answering the four memory-hierarchy questions.

Q1: Where Can a Block Be Placed in a Cache?

Restrictions on where a block is placed create three categories of cache organization:

- If each block has only one place it can appear in the cache, the cache is said to be *direct mapped*. The mapping is usually (block-frame address) modulo (number of blocks in cache).
- If a block can be placed anywhere in the cache, the cache is said to be *fully associative*.

the cache must pick a block to replace; the VAX-11/780 selects one of the two sets at random. Replacing a block means updating the data, the address tag, and the valid bit. Once this is done, the cache goes through a regular hit cycle and returns the data to the CPU.

Writes are more complicated in the VAX-11/780, as they are in any cache. If the word to be written is in the cache, the first four steps are the same. The next step is to write the data in the block, then write the changed-data portion into the

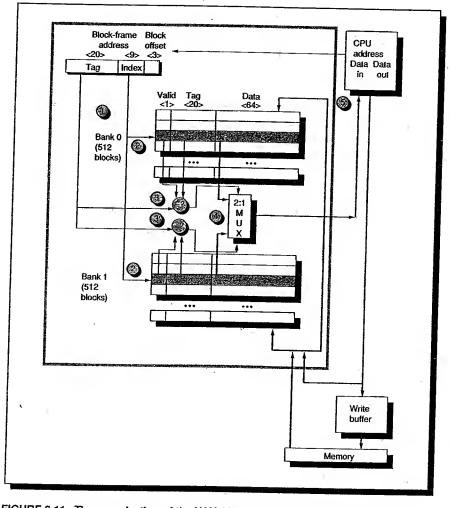

FIGURE 8.11 The organization of the VAX-11/780 cache. The 8-KB cache is two-way set associative with 8-byte blocks. It has 512 sets with two blocks per set; the set is selected by the 9-bit index. The five steps of a read hit, shown as circled numbers in order of occurrence, label this organization. The line from memory to the cache is used on a miss to load the cache. Multiplexing as found in step 4 is not needed in a direct-mapped cache. Note that the offset is connected to chip select of the data SRAMs to allow the proper words to be sent to the 2:1 multiplexer.